DOCUMENT RESUME

ED 074 135 TM 002 498

AUTHOR Williams, Cynthia L.
TITLE Effects of Training on Rating Reliability, as Estimated by ANOVA Procedures, for Fluency Tests of Creativity.
PUB DATE Feb 73
NOTE 31p.; Paper presented at annual meeting of the National Council on Measurement in Education, AERA (New Orleans, Louisiana, February 25-March 1, 1973)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Analysis of Variance; *Evaluation Criteria; *Examiners; *Scoring Formulas; Speeches; *Test Results; *Training Techniques; Volunteer Training
IDENTIFIERS *Divergent Production Battery

ABSTRACT

Each test in the Divergent Production battery requires the examinee to produce a response. Since these responses must be evaluated, the factor of rater judgment influences the reliability of scores. The problem of scoring reliability is one which pervades the literature on creativity research, where either low estimates or no estimates have been reported when tests from the battery are used. The purpose of this study was to develop a training program for raters of some Divergent Production fluency factor tests and to evaluate this program. An experimental design was generated for the evaluation and the scoring reliability was estimated through analysis of variance procedures. General principles for training raters and for analyzing the results of the design will be discussed. (Author)

ED 074135

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

FILMED FROM BEST AVAILABLE COPY

EFFECTS OF TRAINING ON RATING RELIABILITY, AS ESTIMATED BY ANOVA PROCEDURES, FOR FLUENCY TESTS OF CREATIVITY

Cynthia L. Williams
University of Pittsburgh

Presented at the Annual Meeting of the
National Council on Measurement in Education
New Orleans, Louisianna
February, 1973

EFFECTS OF TRAINING ON RATING RELIABILITY, AS ESTIMATED BY ANOVA PROCEDURES, FOR FLUENCY TESTS OF CREATIVITY

Cynthia L. Williams
University of Pittsburgh

Measures of creative mental abilities, such as the divergent production battery developed by Guilford and his colleagues (e.g., Guilford, Wilson, & Christensen, 1952; Guilford, Kettner, & Christensen, 1954), require the examinees to produce a response, given some basic information. Since these responses must then be evaluated, the factor of rater judgement influences the reliability of response scores. Research on the problem of rater judgement has indicated that raters, in general, tend to differ from one another in the scoring criteria applied, to change the scoring criteria for different individuals being rated, and to differ with respect to the distribution of grades throughout the score scale (Coffman, 1971).

The problem of scoring reliability is one which pervades the literature on creativity research, where tests from the divergent production battery are often employed. Many studies which utilize tests from the battery do not report estimates of scoring reliability (Fulgosi & Guilford, 1968; Cropley, 1967; Cline, Richards, & Needham, 1963; Christensen, Guilford, & Wilson, 1957). When scoring reliability estimates are given, they are typically low. For example, Shin (1971) reports scoring reliabilities of 0.81, 0.79, 0.78, 0.74, and 0.68 for two raters of the five divergent production tests used in his study. Curiously, the manuals accompanying the divergent production tests do not contain information regarding scoring reliability. Rather,

alternate form reliability coefficients are reported. Finally, reports of the factor analytic studies on the structure of intellect model from which the divergent production battery was derived do not include scoring reliability estimates, although various internal consistency coefficients are reported (Gershon, Guilford, & Merrifield, 1963; Guilford, Christensen, Frick, & Merrifield, 1957; Guilford, Merrifield, & Cox, 1961; Hoepfner & Guilford, 1965).

As was indicated previously, the lack of consistent scoring criteria across raters produces scoring unreliability. A firmly held belief is that rating errors can be minimized and scoring reliability increased by the careful training of raters (Guilford, 1964). Thus, the major purpose of the research undertaken was twofold. First, procedures for training raters to score protocols from the Utility Test, a test from the divergent production battery, were developed. The Utility Test was selected from the battery of tests available because of its wide use in the creativity research literature. Also, this test provides a measure of ideational fluency, a factor which has been suggested as a prevasive element in the measurement of creativity (Fulgosi & Guilford, 1968; Christensen, Guilford, & Wilson, 1957; Clark & Mirels, 1970; Shin, 1971). Secondly, the procedures were evaluated for their effectiveness in increasing scoring reliability. In addition to the effects of training, the factor of scoring order was investigated. Since the Utility Test contains two parts, one could question whether the scores assigned by raters are a function of the order in which the raters scored each part, that is, scoring Part I first and Part II second as contrasted with scoring Part II first and Part I second. One could also question whether the factor of sequence of scoring systematically

influences the scores assigned. The presence of a sequence effect would indicate that the average score assigned to those protocols scored first differ from the average score assigned to those scored second, regardless of the test part. One final factor investigated was whether the average scores of the two parts of the Utility Test were equal. The investigation of the order, sequence, and test part variables provides information tangential to the major purpose of the study, but allows one to examine potential sources of variation in the general rating situation.

METHOD

A. Development of the Training Procedure

In developing the training procedure, reference was made to other types of measuring devices which use ratings, such as essay examinations and projective techniques, as well as other measures of creativity. Development of consistent scoring standards across all raters appeared to be the major concern of researchers using such devices and several general principles for training raters were identified. The first step in a training program should be one of developing the concept of interest and of establishing a rationale for the measuring procedure. Secondly, the scoring procedure should be made as objective as possible, leaving little room for questions from the raters (Grant & Caplan, 1957). Non-overlapping response categories should be developed and defined precisely. In addition, examples of typical responses occurring in each category should be included. Rater practice in the use of the scoring procedure is a crucial aspect of the training (Tomkins, 1947; Anderson, 1960; Feldt, 1962; Eisner, 1965). In conjunction with these practice sessions, discussions regarding rating discrepancies should be held (Tomkins, 1947;

Eisner, 1965; Feldt, 1962). Finally, Guilford (1964) has suggested that the raters be made aware of the various rating errors, such as leniency errors, relative halo effects, and contrast errors.

Imparting to the rater knowledge about the construct is a primary objective in rater training. This process typically includes a definition of the construct and/or a rationale for the testing procedure. In the manual for the Torrance Tests of Creative Thinking (Torrance, 1966, p. 19), "the importance of familiarity with the rationale of the test tasks and the concepts of fluency, flexibility, originality, and elaboration" is emphasized. Also, included in the scoring guide for this test (pp. 6-16) is a discussion of the rationale for both the figural and verabl tasks. The introduction to the developed training materials contains a brief discussion of divergent production and the structure of intellect model, as proposed by Guilford. The basic factors of fluency, flexibility, orginality, elaboration, redefinition, and sensitivity are briefly defined. Since the function of the program is to train raters to score protocols for fluency, a description of the fluency factor and of some of the proposed measures from the divergent production battery is also included in the introductory sections. Finally, a discussion of scoring reliability and of some sources of rating errors, which can produce scoring unreliability, is provided.

The crucial aspect of the objectification of the scoring procedure is the definition of the scoring categories. The fluency score ascribed to an individual's protocol (a set of responses to a specified task) is the total number of acceptable responses produced. In scoring a protocol for fluency, the rater must classify each particular response as either acceptable or unacceptable. As stated in the technical man-

ual accompanying the Utility Test, an acceptable ideational fluency response has the defining characteristic of relevance (Wilson, Merrifield, & Guilford, 1962). However, the manual for the Plot Titles test, the responses to which can also be scored for ideational fluency, indicates that any response which is relevant, but not a duplicate of a previous response, is acceptable (Berger & Guilford, 1969). While the manuals accompanying the divergent production tests do not explicitly define the terms relevant and duplication, an attempt was made in the developed training materials to define more clearly these two characteristics of an acceptable response.

For the Utility Test, the examinee must write as many uses as he can for a brick (Part I) and for a wooden pencil (Part II). A relevant response in this context must be an example of a possible use for a brick or for a wooden pencil. The critical word is use. In general, use indicates putting some object into service for an intended purpose. The meaning of the word use stresses the practicality of the object for achieving some desired outcome or result. The training materials contain a table of some possible categories of uses for a brick and for a wooden pencil. The suggestion to the raters is to use the list as a device for familiarizing themselves with some types of uses which might be encountered in scoring responses, but not to regard it as a complete listing. In defining the response characteristic of duplication, four situations are described in the training materials.

1. If the response is an exact replication of a previous response, the response occurring second on the list would be a duplicate.

2. Another way in which a response can be a duplicate is if the response is synonymous with a previous response. For example, with reference to

a use of a wooden pencil, "bite it" is essentially synonymous with the response "chew it." If these two responses occurred on a protocol, whichever one occurred second on the list would be a duplicate.

3. Another situation in which duplication occurs is when a response is either a specific or a general case of a previous response. The response which occurs second is a duplicate of the first response. To illustrate this type of duplication, the responses "make a list" and "make a grocery list" can be considered. If the response "make a list" occurred first and "make a grocery list" occurred second on the protocol, then "make a grocery list" is a duplicate, since it is a specific case of the previously given general case "make a list." However, if "make a grocery list" occurred first and "make a list" occurred second, then "make a list" is a duplicate, since it is a general use of a previously given specific case "make a grocery list."

4. The final situation in which duplication occurs is related to the previous situation of general/specific duplication. In this situation what is varied in the response series is the type of specific case. For example, in the series of responses "write a story," "write a poem," "write a speech," "write letters," all of the responses are subsumed under the more general response "to write," However, only the responses subsequent to the first given would be considered duplicates.

Given the definitions of relevance and duplication, the rater must then follow specific rules for classifying a given response as acceptable or unacceptable.

1. If the response is relevant and is not a duplication of a previous response, the response is categorized as acceptable.

2. If the response is relevant, but is also a duplication of a previous

response, the response is categorized as unacceptable.

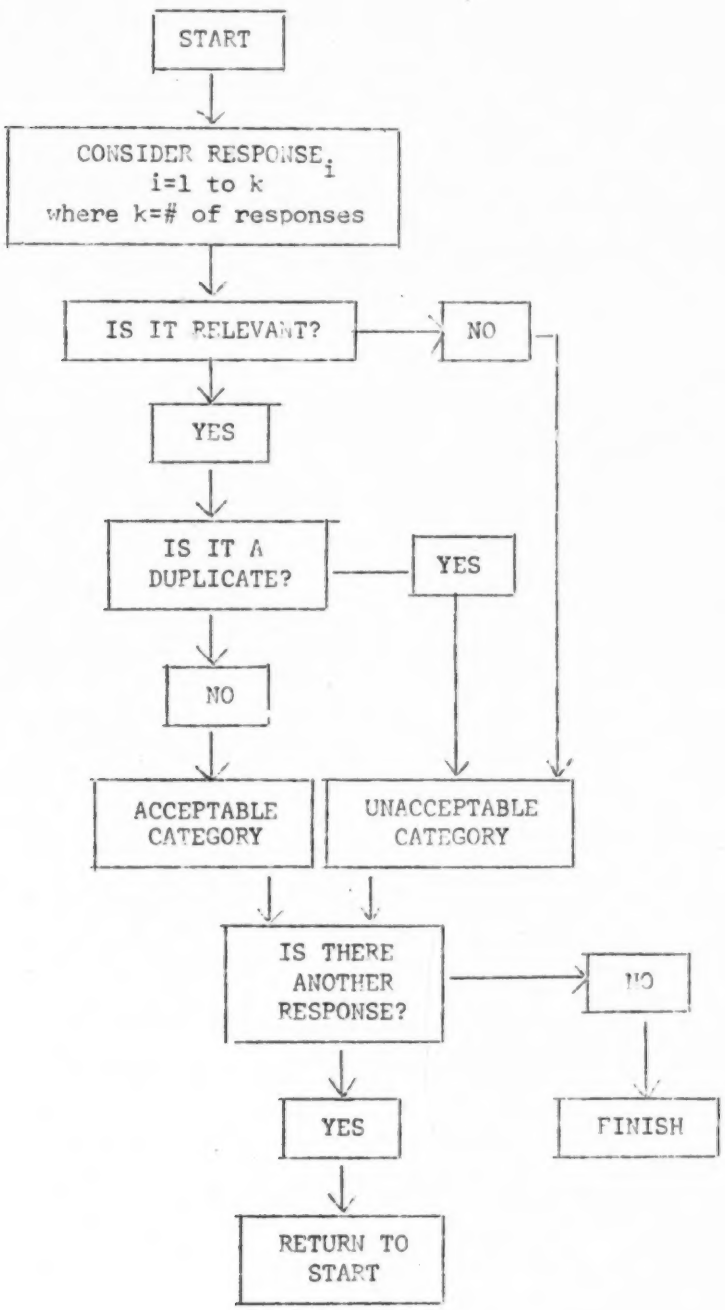
3. If the response is irrelevant, it is automatically categorized as unacceptable.

To illustrate the preceding rules, the process of categorizing the responses is presented in the training materials as a flow chart, which is given in Figure 1. In the categorization of a specific response, the rater must consider two questions. First, does the response provide a relevant example for the task requested? Secondly, if the response is relevant, is it a duplicate of a previous response? When the rater has answered these two questions, utilizing the definitions of relevance and duplication, the response has been categorized as either acceptable or unacceptable.

The final section of the training materials incorporates three suggestions for rater training: provision of examples of acceptable and unacceptable responses, practice in scoring sample responses, and discussions of rating discrepancies. This section of the training materials is structured so that each rater scores the brick responses of three "individuals" and then scores the wooden pencil responses, but the scoring process is done in conjunction with the training manual. The responses of the three "individuals" to both parts of the Utility Test were developed to provide raters with examples of relevant and irrelevant responses and the types of duplication outlined previously. When a rater begins scoring the sample protocols, the instructions in the training materials indicate that he is to consider the first response and decide whether the response is acceptable or unacceptable. In order to compare his decision with a standard, the rater then lifts the slip of paper

FIGURE 2

Procedure Used in Scoring Protocols for Fluency



following the response. Beneath the slip, the correct categorization is given. This process is repeated for each response given on the protocol. With this structure of training materials, all raters can be exposed to the same information, where, if training were conducted with groups of raters, the specific information may be contingent on the nature of the group. Finally, a coding sheet was developed to provide raters with a method for recording their decisions. The coding sheet was designed to include a system for insuring that the number of tallies recorded for the acceptable and unacceptable responses sum to the total number of responses given. A coding sheet of this form should reduce the number of coding errors on the part of the rater. The use of the coding sheet is explained in the training manual and practice in its use is provided during the scoring of the sample protocols.

Prior to the evaluation of the preceding training materials, try-out sessions were conducted. Volunteers were administered the training materials and on completion, independently scored a sample brick protocol and a sample wooden pencil protocol. These protocols were developed to include examples of the relevance and duplication characteristics of responses. The results of the try-out sessions indicated that the transition instructions between the definition of response characteristics and the scoring of sample protocols required clarification. No systematic errors in scoring were indicated by an item by item analysis of rater scoring of the sample protocols. A final observation in the try-out sessions had implications for the evaluation procedure. During the try-out sessions, the raters worked in the same room. Competition between the individuals to finish first or

to keep up with others in the room indicated the importance of raters working independently and in isolated conditions.

B. Evaluation of the Developed Training Procedure

Raters

In most of the research utilizing some portion of the divergent production battery, those responsible for scoring protocols have included the principle researcher and/or members of the staff, graduate or advanced undergraduate students, or teachers involved in the project (for example, Cropley, 1967; Clark & Mirels, 1970; Schmadel, Merrifield, & Bonsall, 1965; Fulgosi & Guilford, 1968; Shin, 1971). To summarize, the general class of raters utilized could be best described as an adult, well-educated, volunteer group. In the present study, volunteers were requested from graduate students in the Department of Educational Research, School of Education, University of Pittsburgh. In the request, students were informed that raters were needed to score responses to a creativity test and that the task should take at most two hours to complete. Information regarding the specific problem and the nature of the variable being considered was withheld. Of the 21 students asked to participate, 20 volunteered their services.

Instrumentation and Protocols

The Utility Test purports to measure the structure of intellect factor of ideational fluency. This test is composed of two parts and in each part the examinee is required to write as many possible uses as he can for a specified object. In Part I the object is a brick and in Part II, a wooden pencil. Five minutes are allotted to each part.

Protocols for the subtests of the divergent production battery,

including the Utility Test, were available from a previous investigation (Shin, 1971). In June, 1971, tests from the divergent production battery were administered to 125 eleventh grade students of a suburban Pittsburgh school district. From this pool of students, 20 individuals were randomly selected. The responses of these 20 individuals to the Utility Test were then reproduced, so that four sets of protocols in the same style of handwriting were available. During the scoring session, each rater received a set of 40 protocols, a set of responses to Part I and to Part II of the Utility Test for 20 individuals.

Training Methods

Two training methods were compared: one labeled the developed training method and the other, the usual training method. With regard to the usual training method, little information about the rater training procedures for divergent production tests is available. Thus, the delineation of the usual training method was derived from an examination of the Utility Test manual. For the purpose of this study, the training procedure referred to as the usual training method consisted of the following procedure. A rater received a package of training materials. Included in this package were general instructions for proceeding through the materials, a copy of the scoring directions provided in the Utility Test manual, and a blank, sample test with a coding sheet for recording scores. The coding sheet provides spaces for recording the names of the individual and the score associated with that individual. The rater was instructed to read the manual and the sample test carefully and to develop for himself a method for recording the total number of acceptable responses given by an individual to each part of the

test. No rationale for the test or the scoring procedure was provided beyond that information included in the materials. Each rater worked independently and was isolated from other raters. No questions specific to the scoring procedure were answered.

Raters trained with the developed training method also received a package of training materials. Included in this package were general instructions for proceeding through the materials and a program which was designed specifically to train raters to score the Utility Test, which was described previously. During the training session, each rater worked independently and was isolated from other raters. Again, no questions specific to the scoring procedures were answered.

Procedure

From the pool of 20 volunteers, ten were randomly selected to be members of the developed training method group. The remaining ten were trained with the usual method. Within the two training groups, five raters were randomly selected to score the responses to Part I first and to Part II second (order 1) and the other five scored protocols in the order Part II first and Part I second (order 2). Each rater was permitted to select the time and the location for participation at his convenience. When a given rater participated, he was provided with a package containing the appropriate training materials. Each rater worked independently and was isolated from other rater who may have also selected to participate at that time. After completing the training session, each rater received the protocols of 20 individuals to score. The instructions for the scoring session indicated to the rater in which order he was to score the protocols. That is, either all responses to Part I were scored first or all responses to Part II were scored first.

No questions regarding the scoring procedure were answered. The order of the 20 individuals was randomized for each rater. Members of the usual training method group averaged approximately one hour in completing both the training and the scoring sessions, while members of the developed training method group averaged approximately two hours in completing both sessions. Within a period of one week, all raters had participated in the study.

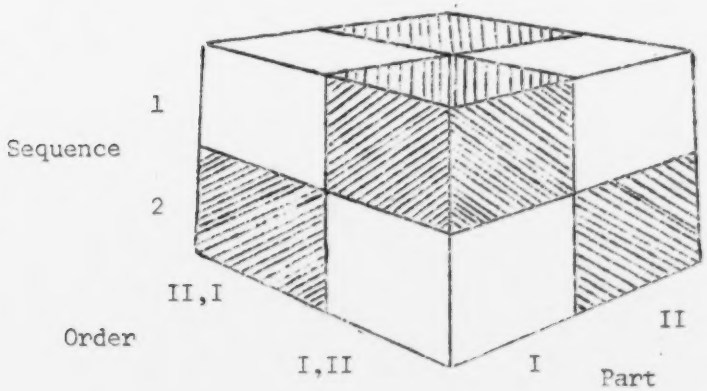
ANALYSIS

A. Design

In the present study, six main sources of variation were investigated: type of training procedure, raters, individuals, test part, sequence in scoring, and order of scoring each test part. In each of the two training procedures, one half of the raters (5 R) scored the responses of 20 Individuals (I) to Part I first and to Part II second (order I, II), while the remaining half of the raters scored Part II first and Part I second (order II, I). In addition, the scores were assigned over a sequence factor: scores assigned first (sequence 1) and scores assigned second (sequence 2). Considering the part, order, and sequence variables, a 2 X 2 X 2 factorial design with eight design cells can be generated, as shown in Figure 2. However, when the nature of each cell is investigated, certain cell combinations do not exist. Given the order I, II, raters can not possibily score Part I first and Part II second. Similarly with the order II, I, the cells corresponding to raters scoring Part I in sequence 1 and Part II in sequence 2 do not exist. Those cross-hatched cells in Figure 2 represent those conditions which exist for the present study. If the rater and individual dimensions are added to the existing cell combinations indicated above,

FIGURE 2

Representation of the Present Study



the design can be represented as in Figure 3. To add the training method variable to the design in Figure 3 would duplicate that design so that two such designs exist, one for the developed training method and one for the usual training method. While the individual dimension crosses all of the factor levels, raters are nested within order and training method, but are crossed with the part, sequence, and individual variables.

FIGURE 3

Design of the Present Study, Not Including Training Dimension

						Rater 1			Rater 2			...	Rater 5		
						Ind'ual			Ind'ual				Ind'ual		
						1	...	20	1	...	20	...	1	...	20
Order	I,II	Part	I	Sequence	1										
			II		2										

						Rater 6			...	Rater 10		
						Ind'ual				Ind'ual		
						1	...	20	...	1	...	20
Order	II,I	Part	II	Sequence	1							
			I		2							

The resulting designs is best described as a fractional hierachical design and, thus, many of the sources of variation are confounded (Cox, 1958, pp. 247-268; Kirk, 1968, pp. 385-387). Confounding in a design means that some, or all, of the sources of variation can not be separated, logically or mathematically, from other sources of variation in the design. Table 1 presents the sources of variation, the alias, or confounded, terms of the design, and the expected mean squares. The training method is indicated with the letter T; individuals, with I: raters which are nested within orders and training method, with R (OT): scoring order, with O; test part, with P; and sequence, with S. The remaining terms are the appropriate interaction terms. Nesting factors are placed in parentheses. In addition to indicating the sources of variation, the capital letters have also been used in the specification of the coefficients for the expected mean squares. For example, the coefficient of the variance component σ^2_{IT} is LPSR, where LPSR equals the product of the number of order levels (L), the number of test parts (P), the number of sequence levels (S), and the number of raters within a given order and training method (R). In stating the linear model for the data, Kirk (1968, p. 390) suggests a notation which includes the alias terms. Thus, the model to be analyzed can be stated as

$$
\begin{aligned}
Y_{iropst} = \mu + a_i + b_{r(ot)} + \gamma_o \{\delta\pi_{ps}\} + \delta_p \{\gamma\pi_{os}\} + \pi_s \{\gamma\delta_{op}\} \\
+ \tau_t + ab_{ir(ot)} + a\gamma_{io} \{a\delta\pi_{ips}\} + ac_{ip} \{a\gamma\pi_{ios}\} \\
+ a\pi_{is} \{a\gamma\delta_{iop}\} + a\tau_{it} + \delta b_{pr(ot)} \{\pi b_{sr(ot)}\} \\
+ \tau\gamma_{to} \{\tau\delta\pi_{tps}\} + \tau\delta_{tp} \{\tau\gamma\pi_{tos}\} + \tau\pi_{ts} \{\tau\gamma\delta_{top}\} \\
+ \tau a\delta_{tio} \{\tau a\delta\pi_{tips}\} + \tau a\delta_{tip} \{\tau a\delta\pi_{tios}\} \\
+ \tau a\pi_{tis} \{\tau a\gamma\delta_{tiop}\} + \delta ab_{pir(ot)} \{\pi ab_{sir(ot)}\} + e_{(iropst)}.
\end{aligned}
\qquad (1)
$$

TABLE 1

Sources of Variation, Alias Terms, and Expected Mean Squares of the Study

Source	Alias	Expected Mean Square
I		$\sigma^2 + LPSRT\ \sigma^2_{IT} + PS\ \sigma^2_{IR(OT)} + LPSRT\ \sigma^2_{I}$
R(OT)		$\sigma^2 + PS\ \sigma^2_{IR(OT)} + PSI\ \sigma^2_{R(OT)}$
O	PS	$\sigma^2 + PSRT\ \sigma^2_{IO} + PS\ \sigma^2_{IR(OT)} + PSI\ \sigma^2_{R(OT)} + PSIRT\ \sigma^2_{O}$
P	OS	$\sigma^2 + S\ \sigma^2_{PIR(OT)} + SI\ \sigma^2_{PR(OT)} + LSRT\ \sigma^2_{IP} + LSIRT\ \sigma^2_{P}$
S	OP	$\sigma^2 + P\ \sigma^2_{PIR(OT)} + PI\ \sigma^2_{PR(OT)} + LPRT\ \sigma^2_{IS} + LPIRT\ \sigma^2_{S}$
T		$\sigma^2 + LPSR\ \sigma^2_{IT} + PS\ \sigma^2_{IR(OT)} + PSI\ \sigma^2_{R(OT)} + LPSIR\ \sigma^2_{T}$
IR(OT)		$\sigma^2 + PS\ \sigma^2_{IR(OT)}$
IO	IPS	$\sigma^2 + PS\ \sigma^2_{IR(OT)} + PSRT\ \sigma^2_{IO}$
IP	IOS	$\sigma^2 + S\ \sigma^2_{PIR(OT)} + LSRT\ \sigma^2_{IP}$
IS	IOP	$\sigma^2 + P\ \sigma^2_{PIR(OT)} + LPRT\ \sigma^2_{IS}$
IT		$\sigma^2 + LPSR\ \sigma^2_{IT}$
PR(OT)	SR(OT)	$\sigma^2 + S\ \sigma^2_{PIR(OT)} + SI\ \sigma^2_{PR(OT)}$
TO	TPS	$\sigma^2 + PSR\ \sigma^2_{TIO} + PS\ \sigma^2_{IR(OT)} + PSI\ \sigma^2_{R(OT)} + PSIR\ \sigma^2_{TO}$
TP	TOS	$\sigma^2 + S\ \sigma^2_{PIR(OT)} + LSR\ \sigma^2_{TIP} + SI\ \sigma^2_{PR(OT)} + LSIR\ \sigma^2_{TP}$
TS	TOP	$\sigma^2 + P\ \sigma^2_{PIR(OT)} + LPR\ \sigma^2_{TIS} + PI\ \sigma^2_{PR(OT)} + PSR\ \sigma^2_{TIO}$
TIO	TIPS	$\sigma^2 + PS\ \sigma^2_{IR(OT)} + PSR\ \sigma^2_{TIO}$
TIP	TIOS	$\sigma^2 + S\ \sigma^2_{PIR(OT)} + LSR\ \sigma^2_{TIP}$
TIS	TIOP	$\sigma^2 + P\ \sigma^2_{PIR(OT)} + LPR\ \sigma^2_{TIS}$
PIR(OT)	SIR(OT)	$\sigma^2 + S\ \sigma^2_{PIR(OT)}$

Random effects (individuals and raters) are represented by Latin letters, while fixed effects are indicated by Greek letters. However, any interaction term containing a random factor is also a random factor. The terms within the braces are the aliases of the corresponding sources of variation.

B. Estimation of Scoring Reliability

When the theoretical definition of reliability, the ratio of the true score variance to the observed score variance, is recalled, the problem of partitioning the observed score variance into its true score and error score components becomes evident. Burt (1955) succinctly pointed out the problem when he remarked,

> We have seen that a reliability coefficient is intended to indicate the ratio of the estimated variance of the "true" measurements to the actual variance of the observed measurements, i.e., to the total variance" conceived as the sum of the "true variance" and an "error variance." But how do we know that the value taken in the numerator in the ratio just calculated really represents the "true variance" we have in mind, and that it does not incorporate something that we might (if we knew its real nature) also might regard as error? (p. 115)

Estimating reliability by correlational methods does not permit the investigator to partition the observed score variance, except at a gross level. The analysis of variance, on the other hand, allows the possibility for such partitioning of the observed score, of total variance. Through the use of experimental design and the analysis of variance, factors which affect the reliability estimates can be identified and more precise estimates of reliability can be obtained. The use of analysis of variance procedures to estimate test reliability, in general, and ratings, specifically, has been suggested (Hoyt, 1941; Ebel, 1951). Typically, two sources of variation are identified: individuals and

test items (test reliability) or individuals and test raters (rating reliability). However, the principles of design can be extended to more complex designs, where a number of variables can be investigated (Stanley, 1962).

To estimate the scoring reliability for each training procedures, the sums of squares and mean squares were computed separately for each group. The model analyzed can be derived from formula (1) by excluding from the model any source of variation which contains the training method and eliminating the training method as a nesting variable. The expected mean squares for this derived model can be obtained from Table 1 in a similar manner. Table 2 presents the summary table used in estimating the scoring reliability.

Given the true and error score model for estimating the average reliability of ratings with the data analyzed in the analysis of variance procedure. The general formula is given by

$$\text{avg } r = \frac{MS_s - MS_e}{MS_s + (k-1)\, MS_e} \qquad (3)$$

where MS_s is the mean square for subjects, MS_e is the error mean square, and k is the number of raters. To estimate the average scoring reliabilities of the two training groups, formula (2) was restated in terms of the present design. The appropriate error mean square for the individual mean square (MS_I) is the mean square for the individual by rater interaction (MS_{IR}) and since five raters are nested within the two orders, (k-1) is equal to eight. Substituting the appropriate values into formula (2), a scoring reliability of 0.92432 was obtained for the raters trained with the developed materials, while a scoring relia-

TABLE 2

Summary Table for Reliability Estimates

Source[1]	df	SS_{DTM}[2]	MS_{DTM}	SS_{UTM}[3]	MS_{UTM}
I	19	4205.90	221.3632	6751.4475	355.3393
R(O)	8	90.66	11.3325	2591.2600	323.9075
IR(O)	152	303.34	1.9957	1140.1400	7.5009
O {PS}	1	1.44	1.4400	85.5625	85.5625
P {OS}	1	294.64	294.6400	578.4025	578.4025
S {OP}	1	9.00	9.0000	6.0025	6.0025
IO {IPS}	19	27.66	1.4588	341.9875	17.9993
IP {IOS}	19	730.06	38.4242	1901.3475	100.0709
IS {IOP}	19	42.10	2.2158	75.7475	3.3795
OR(O) {SR(O)}	8	93.76	11.7200	89.3200	11.1650
PIR(O) {SIR(O)}	152	305.44	2.0095	513.6800	3.7947

[1]The aliases of the sources of variation are in braces.
[2]DTM indicates the developed training method.
[3]UTM indicates the usual training method.

bility of 0.83755 was obtained with the usual training procedure. Ebel (1951) showed that formula (2) is equivalent to the average intercorrelation between all possible pairs of raters. Thus, these reliability estimates can be thought of as <u>average</u> estimates of rating reliability for each training group.

Examining the components in the expected mean square, Ebel (1951) showed that one could either include the "between-raters" variance in the formula or could exclude the term. The inclusion or exclusion of the term, however, depends upon how the ratings are to be used.

> Specifically, the "between-raters" variance should be removed where the final ratings on which decisions are based consist of averages of complete sets of ratings from all observers, or ratings which have been equated from rater to rater such as ranks, Z-scores, etc. Likewise, if comparisons are never made practically, but only experimentally, between ratings of pupils by different raters, the "between-raters" variance should be removed. But if decisions are made in practice by comparing single "raw" scores assigned to different pupils by different raters, or by comparing averages which came from different groups of raters, then the "between-raters" variance should be included as part of the error terms. (p.412)

When the "between"raters" variance is included in the error term, the following formula (Ebel, 1951) is appropriate for estimating scoring reliability:

$$\text{avg "br" } r = \frac{SS_s - \frac{(SS_c + SS_e)}{k - 1}}{SS_s + (SS_c + SS_e)} \qquad (3)$$

where SS_s is the sum of squares for subjects, SS_c is the sum of squares for raters, SS_e is the error sum of squares, and k is the number of raters. Again, substituting the appropriate sums of squares and the value eight for (k-1) into formula (3), the scoring reliability estimate which includes the "between-raters" variance is 0.90364 for the developed training method group and 0.62674 for the usual training group. Formula (3) also provides an average estimate of rating reliability, except that the variability between raters is included in the error term. Thus, both formulas (2) and (3) are equivalent to the average intercorrelation between all pairs of raters of the between raters variance is either included in the error term for both computations or excluded in both computations.

However, if one is interested in the reliability of the average of ratings, Ebel (1951) has shown that this reliability estimate is

equivalent to

$$r_{avg} = \frac{MS_s - MS_e}{MS_e} \qquad (4)$$

where MS_s and MS_e are interpreted as in formula (2). This formula (4) can be obtained through the application of the Spearman-Brown formula to formula (2). If one were to find the average rating assigned to each individual and then to obtain a reliability estimate of these averages, this estimate would be equivalent to that obtained using formula (4). A reliability estimate of the average rating is 0.99098 for the developed training method and 0.97889 for the usual training method.

When the three reliability estimates of the two rater training groups are compared, two results become evident. First, the reliability estimates for the developed training method group are consistently higher than those obtained for the usual training method group. In addition, the reliability estimates for the developed training method group are greater than 0.90, while in the usual method group, only the estimate for the reliability of the average score is greater than 0.90. The second factor to consider is a comparison between the two training groups of the average reliability estimate when the between rater variance is excluded and when it is included. In general, when the between rater variance is included, the reliability estimate is lower than when the estimate excludes this variance. This reduction in reliability is reasonable, since more variability is being included into the error component. For the present study, when the between rater variance is included, the reliability estimate is reduced from 0.83754 to 0.62674 in the usual training method, while the reduction in the estimate for the developed training method is from 0.92432 to 0.903639.

Typcially, the scoring reliability estimates reported are based on the use of computed formula (2). This differential reduction in the scoring reliability estimates serves to indicate the presence of more variability among the raters who were trained with the usual procedure than among those trained with the developed procedure.

C. Analysis of Training Design

The results of the training design are presented in Table 3. As shown in Table 1, no appropriate mean squares for the denominator of an F ratio are available for some sources of variation. For example, no other source of variation has the expected mean square of σ^2 + PSRT σ^2_{IO} + PS $\sigma^2_{IR(OT)}$ + PSI $\sigma^2_{R(OT)}$, which would provide the appropriate denominator of an F ratio to test the O source of variation. In order to test those sources for which no appropriate denominator was available, quasi-F ratios (Kirk, 1968, pp. 212-214) were formed and the degrees of freedom for these ratios were computed, as indicated in Table 3. Since the design contains only one observation per cell, no direct estimate of the within cell variability is available. However, the tests for the I, T, IR(OT), and IT sources of variation required an estimate of this variability. Therefore, the highest order interaction, PIR(OT), was assumed to be zero and the mean square associated with this interaction was used as the estimate of the within cell variability. To summarize the results shown in Table 3, three areas of interest can be identified: the effects of training, the effects of the rating situation, and the effects associated with the test part.

Estimating the mean score assigned in each of the two training groups, the mean score of the experimental group is estimated as 7.35 and the mean score of the control group is 13.0725. The significant

TABLE 3

ANOVA Summary Table for the Training Design

Source	df	SS	MS	F	$df_1 df_2$	p
I	19	9424.02375	496.00125	5.83614[1]	19,21[2]	<0.001
R(OT)	16	2681.92000	167.62000	35.30113	16,304	<0.001
O{PS}	1	32.40125	32.40125	.21004[1]	1,18[2]	0.652
P{OS}	1	794.01125	794.01125	49.20734[1]	1,31[2]	<0.001
S{OP}	1	.15125	.15125	.20005[1]	173,23[2]	>0.999
T	1	6549.40125	6549.40125	26.38556[1]	1,29[2]	<0.001
IR(OT)	304	1443.48000	4.74829	1.76224	304,304	<0.001
IO{IPS}	19	178.82375	9.41178	1.98214	19,304	0.009
IP{IOS}	19	1995.51375	105.02704	38.97874	19,304	<0.001
IS{IOP}	19	52.87375	2.78283	1.03279	19,304	0.423
IT	19	1533.32375	80.70125	29.95069	19,304	<0.001
PR(OT){SR(OT)}	16	183.08000	11.44250	4.26666	16,304	<0.001
TO{TPS}	1	54.60125	54.60125	.33406[1]	1,18[2]	0.570
TP{TOS}	1	34.03125	34.03125	.81775[1]	1,30[2]	0.373
T[TOP]	1	14.85125	14.85125	.81457[1]	1,25[2]	0.375
TIO{TIPS}	19	190.82375	10.04336	2.11515	19,304	0.005
TIP{TIOS}	19	635.89375	33.46809	12.48783	19,304	<0.001
TIS{TIOP}	19	64.97395	3.41967	1.26914	19,304	0.202
PIR(OT){SIR(OT)}	304	819.12000	2.69447			

[1]Quasi-F ratio of the general form $\frac{MS_1 + MS_4}{MS_2 + MS_3}$

[2]Degrees of freedom for the numerator and denominator, respectively, of the quasi-F ratio

difference between the mean score assigned in the two training groups is most likely associated with the definition of fluency developed in the training materials. The definition is more precise and, with the inclusion of the response characteristic of duplication, more restrictive than the definition provided in the Utility Test manual. However, given the discrepancies in the definition of ideational fluency across tests which support to measure this factor, the restriction provided in the developed training materials is appropriate. In addition, the presence of the significant IT, TIP, and TIO interactions indicates that the training procedure has a differential effect in conjunction with different individuals and combinations of test part and scoring order. The nature of confounding, however, complicates this analysis. Specifically, the fact that the TIP and TIO sources are aliased with the TIOS and TIPS sources respectively, does not allow for a precise interpretation of the interactions. To analyze these interactive effects would require further investigations with unconfounded designs.

The results of the present study also provide insights into the general rating situation. The presence of a significant individual effect indicates variability among the sources associated with individuals in the population and an estimate of this variability, $\hat{\sigma}_I^2$, is 5.1656. Guilford (1964) has termed this variability in the rating situation absolute halo, which reflects true variation among individuals being rated. The presence of the significant rater effect indicates variability among the raters and an estimate, $\hat{\sigma}_R^2$, is 2.0359. Rater variability has been termed leniency error by Guilford (1964) and can be interpreted as systematic differences between raters in the scores assigned. The

estimates of the average scoring reliability which includes rater variance would indicate that more rater variability is associated with raters trained with the usual method than those trained with the developed method. On the basis of this information one can conjecture that the developed training procedure reduces the between rater variability.

In addition to the significant individual and rater main effects, two interactions involving the rater dimension were also significant: IR(OT) and PR(OT). The IR(OT) interaction indicates the tendency for raters to rate individuals differentially and has been termed the relative halo effect (Guilford, 1964). Further investigation of this relative halo effect would be facilitated by including the level of rater creativity into the design. In other words, the tendency of raters to rate individuals differentially may be a function of the rater's own capacity. Similarly, the PR (OT) interaction indicates the tendency for raters to score the test parts differentially and provides evidence for the presence of a contrast rating error (Guilford, 1964). However, this interaction is confounded with the SR(OT) interaction, so that the interpretation of the PR(OT) as reflecting a contrast error is only tentative.

Finally, a significant difference between the scores assigned to the test parts is evidenced. An estimate of the mean associated with the brick task is 11.2075 and of the mean associated with the wooden pencil task, 9.2150. Although the part main effect is confounded with the order/sequence interaction, the presence of this difference should necessitate the reconsideration of the task equivalency. Researchers

have assumed the equivalency of the brick and wooden pencil tasks. However, this equivalency is questionable in light of the significant P main effect and requires additional investigation. The presence of a significant IP interaction, which is confounded with the IOS interaction, would indicate the differential response of individuals to the test parts. Again, this interpretation is contingent upon the presence or absence of the IOS interaction.

CONCLUSIONS

In summary, the two major effects of the developed training procedure were to maintain scoring reliability at a level greater than 0.90 and to reduce the average fluency score assigned. As pointed out previously, the estimates of the scoring reliability for raters trained with the developed training method are consistently higher than for those trained with the usual method. In addition, the scoring reliability estimates are maintained at a level which is of practical significance in the further use of the Utility Test. This level of scoring reliability is maintained even when the between rater variance is included in the reliability estimate. Such results are not obtained when the scoring reliabilities for raters trained with the usual procedure are estimated. These results strongly suggest the effectiveness of training raters to score protocols for ideational fluency.

The training procedure developed for the present study consisted of a number of conponents: the inclusion of the rationale and theory of the Utility Test, the definition of the response characteristics of relevance and duplication, scoring practice, and discussion involving scoring discrepancies. To indicate which factor in the training procedure produced the results would require additional investigations.

However, one can consider the generalizability of the training model utilized in the present study to the other measures of creativity available, as well as to the other factors of creativity. The present study has involved only the factor of ideational fluency: the factors of flexibility, elaboration, and orginality and the associated tests were not considered. To apply the training procedure model would necessitate a careful and precise delineation of the factors involved. This definitional process is highly recommended. Unless the raters are provided with clear guidelines for scoring the factors and with practice, one would expect rater variability to be greater than when raters are provided with such information. In other words, the definition of the scoring categories and practice in using these categories is seen as an integral part of training.

Many issues in the area of creativity research are still unresolved. For example, the establishment of creative mental abilities as a construct distinct from that of intelligence has not been confirmed (McNemar, 1964). Also, the relationships between intelligence, creative mental abilities, and academic achievement (Shin, 1971) are yet unclear. Given the evidence for scoring unreliability, one can conjecture that the relationships and research inconsistencies might be made more definitive, if scoring reliability were improved.

REFERENCES

Anderson, C. The new STEP essay test as a measure of composition ability. Educational and Psychological Measurement, 1960, 20, 95-102.

Berger, R. & Guilford, J.P. Plot Titles. Beverly Hills, California: Sheridan Psychological Services, Inc., 1969.

Burt, C. Test reliability estimated by analysis of variance. British Journal of Statistical Psychology, 1955, 8, 103-118.

Christensen, P.R., Guilford, J.P., & Wilson, R.C. Relations of creative responses to working time and instructions. Journal of Experimental Psychology, 1957, 53, 82-89.

Clark, P.M. & Mirels, H.L. Fluency as a pervasive element in the measurement of creativity. Journal of Educational Measurement, 1970, 7 (Summer), 83-86.

Cline, V., Richards, J. & Needham, W. Creativity tests and achievement in high school science. Journal of Applied Psychology, 1963, 47, 184-189.

Coffman, W.E. Essay examinations. In R.L. Thorndike (Ed.), Educational Measurement. Washington, D.C.: American Council on Education, 1971.

Cox, D.R. Planning of experiments. New York: John Wiley & Sons, Inc., 1958.

Cropley, J.A. Creativity and intelligence. British Journal of Educational Psychology, 1966, 36, 259-265.

Ebel, R.L. Estimation of the reliability of ratings. Psychometrika, 1951, 16, 407-424.

Eisner, E. Research in creativity: Some findings and conceptions. Childhood Education, 1963, 39 (April), 371-375.

Feldt, L.S. The reliability of measures of handwriting ability. Journal of Educational Psychology, 1962, 53, 288-292.

Fulgosi, A. & Guilford, J.P. Short term incubation in divergent production. American Journal of Psychology, 1968, 81, (2), 241-246.

Gershon, A., Guilford, J.P., & Merrifield, P.R. Figural and symbolic abilities in adolescent and adult populations. Reports from the Psychological Laboratory. Los Angeles: University of Southern California, 1963, No. 29.

Grant, D.L. & Caplan, N. Studies in the reliability of the short answer essay examination. Journal of Educational Research, 1957, 51, 109-116.

Guilford, J.P. Psychometric methods. New York: McGraw-Hill, 1964.

Guilford, J.P., Christensen, P.R., Frick, J.W., & Merrifield, P.R. The relations of creative thinking to nonaptitude personality traits. Reports from the Psychological Laboratory. Los Angeles: University of Southern California, 1957, No. 20.

Guilford, J.P., Kettner, N.W., & Christensen, P.R. A factor analytic study across the domains of reasoning, creativity, and evaluation, I. Hypotheses and description of tests. Reports from the Psychological Laboratory. Los Angeles: University of Southern California, 1954, No. 11.

Guilford, J.P. Merrifield, P.R., & Cox, A.B. Creative thinking in children at the junior high school levels. Reports from the Psychological Laboratory. Los Angeles: University of Southern California, 1961, No. 26.

Guilford, J.P., Wilson, R.C., & Christensen, P.R. A factor analytic study of creative thinking: II. Administration of tests and analysis of results. Reports from the Psychological Laboratory. Los Angeles, University of Southern California, 1952, No. 8.

Hoepfner, R. & Guilford. Figural, symbolic, and semantic factors of creative potential in ninth-grade students. Reports from the Psychological Laboratory. Los Angeles: University of Southern California, 1965, No. 35.

Hoyt, C. Test reliability estimated by analysis of variance. Psychometrika, 1941, 6, 153-160.

Kirk, R. Experimental design: Procedures for the behavioral sciences. Belmont, Ca.: Brooks/Cole, 1968.

McNemar, Q. Lost: Our intelligence. Why? American Psychologist, 1964, 19, 871-882.

Schmadel, E., Merrifield, P.R., & Bonsall, M. Comparison of performance of gifted and non-gifted children on selected measures of creativity. California Journal of Educational Research, 1965, 16 (May), 123-128.

Shin, Se Ho. Creativity, intelligence, and achievement: A study of the relationship between creativity and intelligence, and their effects upon achievement. Unpublished doctoral dissertation, University of Pittsburgh, 1971.

Stanley, J.C. Analysis of variance principles applied to the grading of essay tests. Journal of Experimental Education, 1962, 30, 279-283.

Tomkins, S.S. The thematic apperception test. New York: Grune & Stratton, 1947.

Torrance, E.P. Torrance tests of creative thinking. Princeton, N.J.: Personnel Press, 1966.

Wilson, R.C., Merrifield, P.R., & Guilford, J.P. Utility Test. Beverly Hills, California: Sheridan Supply Co., 1962.